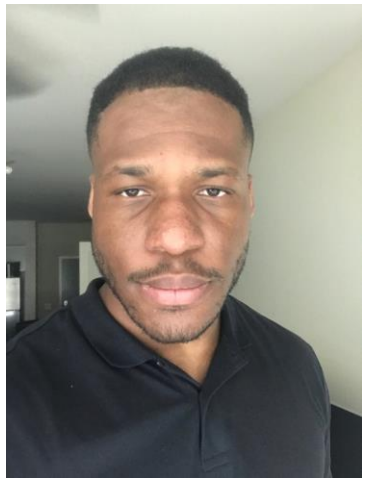
RIFLE: **R**obust **I**nference and Imputation **F**rom **L**ow Ord**e**r Marginals

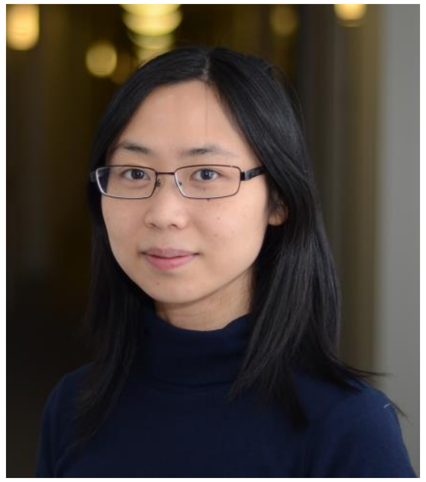
Sina Baharlouei

Daniel J. Epstein Department of Industrial and Systems Engineering



Kelechi Ogudu

Peng Dai



Sze-chuan Suen

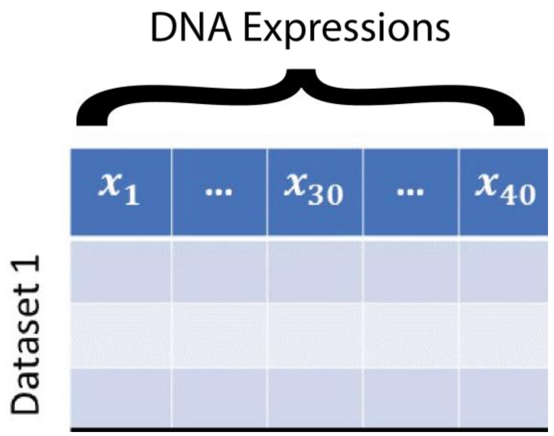
Meisam Razaviyayn

Prevalence of Missing Values in Data-driven Tasks

- Blank answers in questionnaires
- Limitations of data gathering

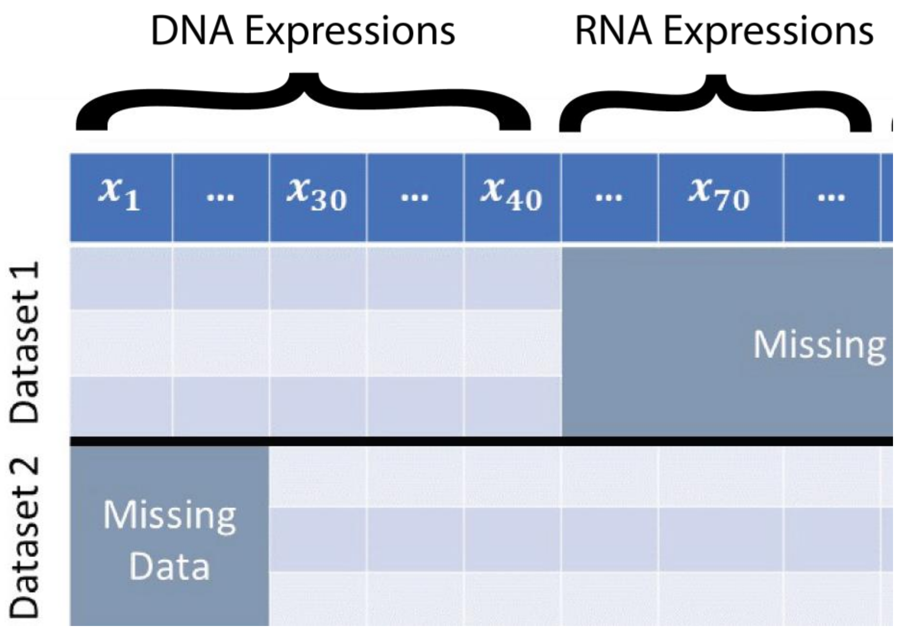
Prevalence of Missing Values in Data-driven Tasks

- Blank answers in questionnaires
- Limitations of data gathering



Prevalence of Missing Values in Data-driven Tasks

- Blank answers in questionnaires
- Limitations of data gathering



Prevalence of Missing Values in Data-driven Tasks

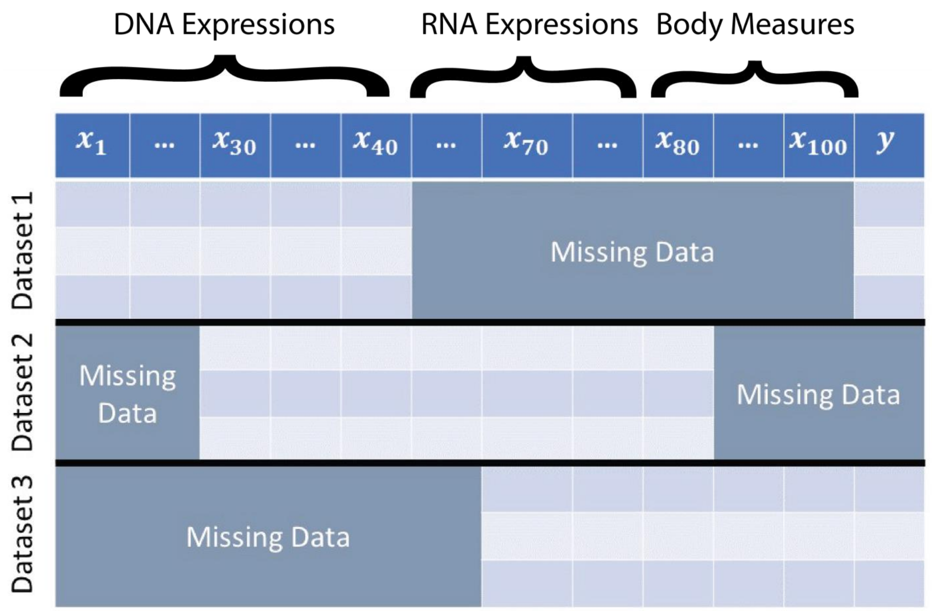
- Blank answers in questionnaires

- Limitations of data gathering

- Blocks of missing values after merging different datasets
  - Related studies from different labs

Existing Approaches for Supervised Learning in the Presence of Missing Data

- **Removing** the rows containing missing entries
  - Losing information

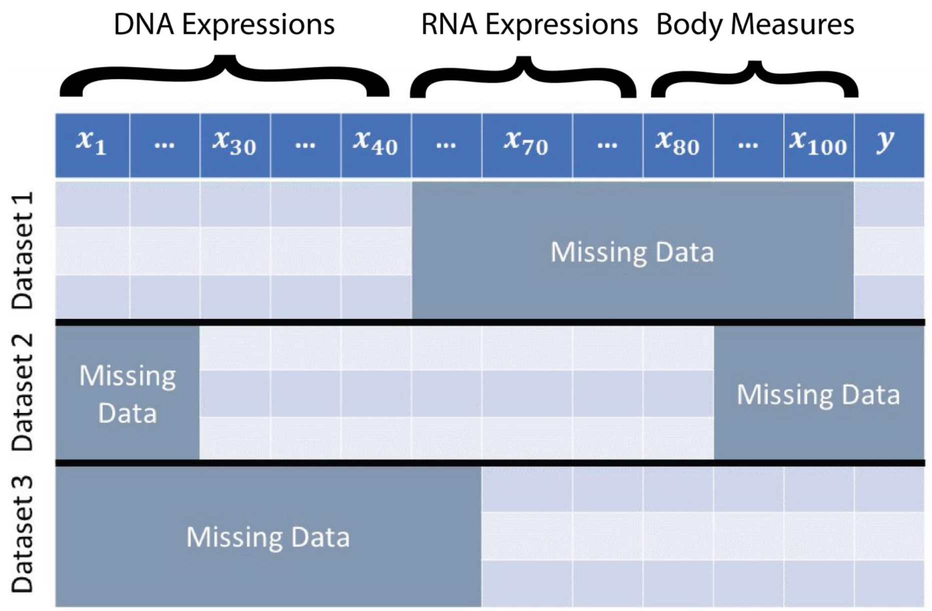


Existing Approaches for Supervised Learning in the Presence of Missing Data

- **Removing** the rows containing missing entries
  - Losing information
- **Imputation and then prediction**
  - Mean/Median imputation
  - Expectation Maximization [Little and Rubin, 1977]
  - KNN Imputer [Troyanskaya et al., 2001]
  - MissForest [Stekhoven et al., 2012]
  - Generative Adversarial Imputation Nets (GAIN) [Yoon et al., 2018]
  - **The imputation error propagates to the prediction phase**

Existing Approaches for Supervised Learning in the Presence of Missing Data

- **Removing** the rows containing missing entries
  - Losing information



- **Imputation and then prediction**
  - Mean/Median imputation
  - Expectation Maximization [Little and Rubin, 1977]
  - KNN Imputer [Troyanskaya et al., 2001]
  - MissForest [Stekhoven et al., 2012]
  - Generative Adversarial Imputation Nets (GAIN) [Yoon et al., 2018]
  - **The imputation error propagates to the prediction phase**

- Prediction without imputation
  - Robust Optimization over uncertainty sets

Robust Inference without Imputation

- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

Robust Inference without Imputation

- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

Robust Inference without Imputation

- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data

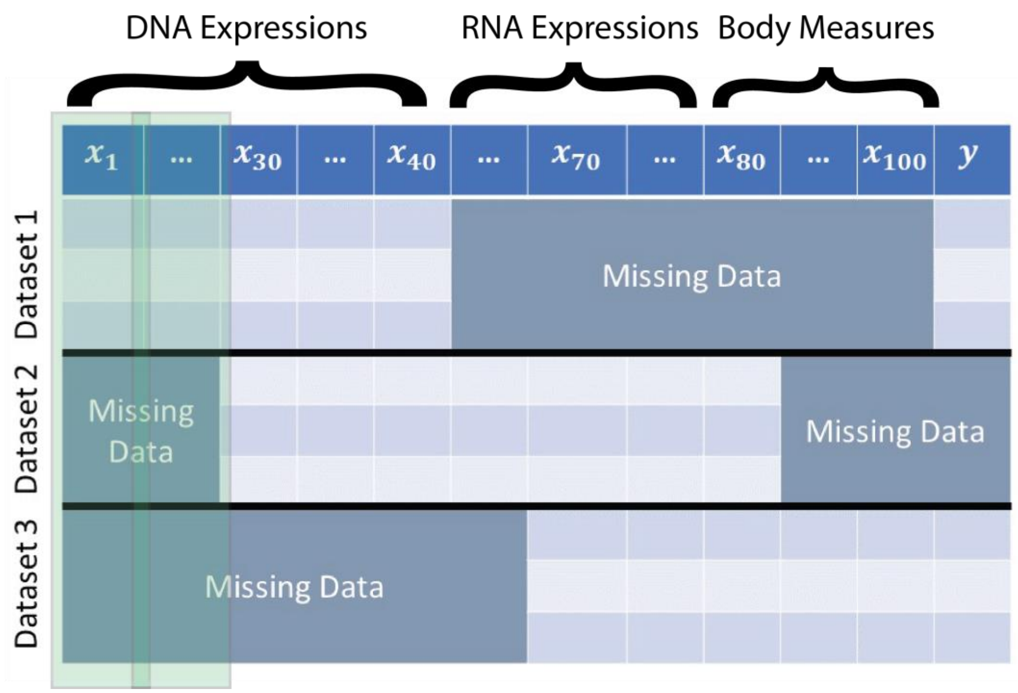
Robust Inference without Imputation

- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data



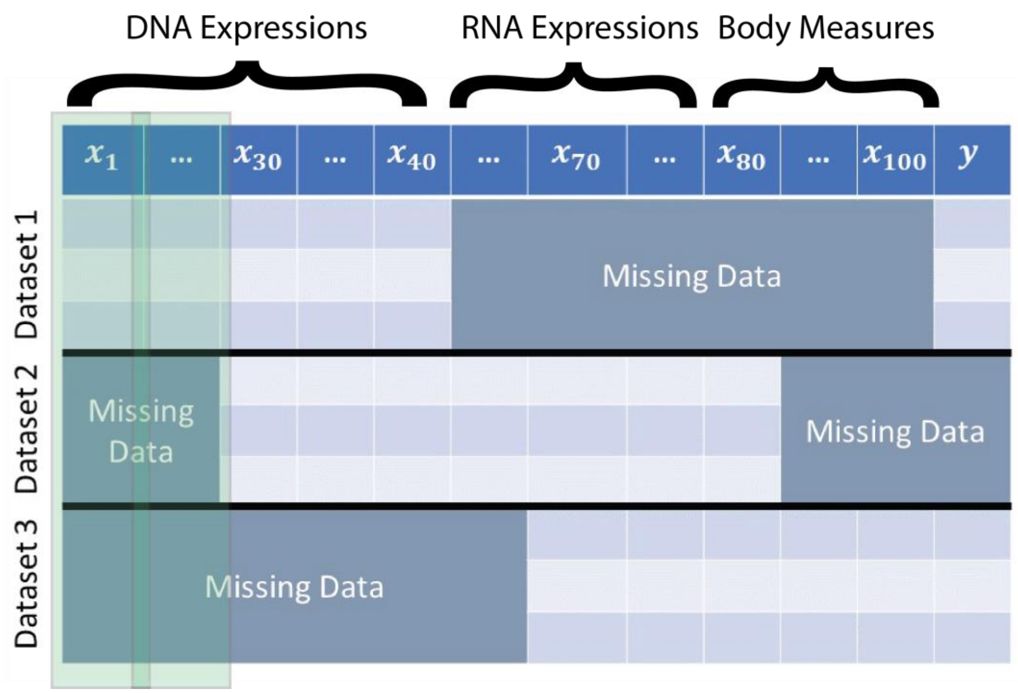
Robust Inference without Imputation

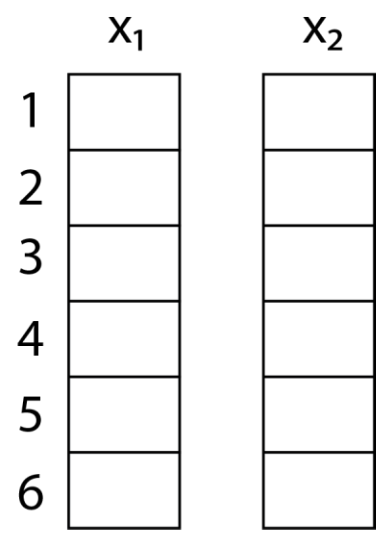
- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data





$$\mathbb{E}[x_1^T x_2] \approx \frac{1}{6}\sum_{i=1}^{6} x_{1i}x_{2i}$$

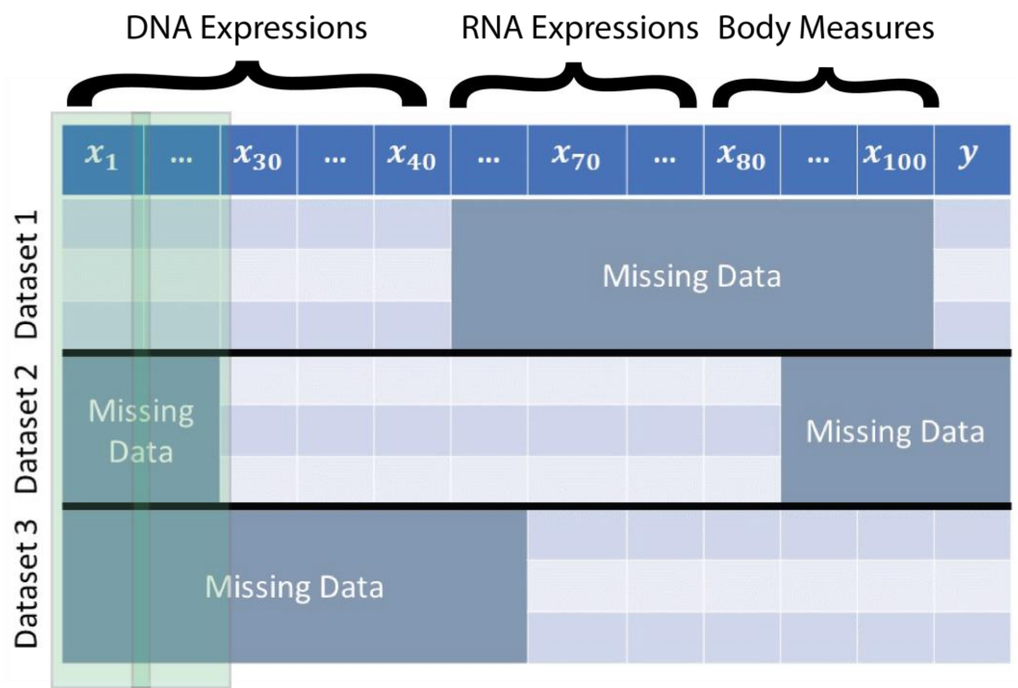
Robust Inference without Imputation

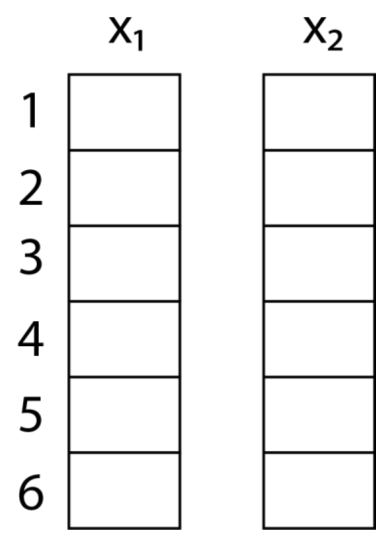
- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

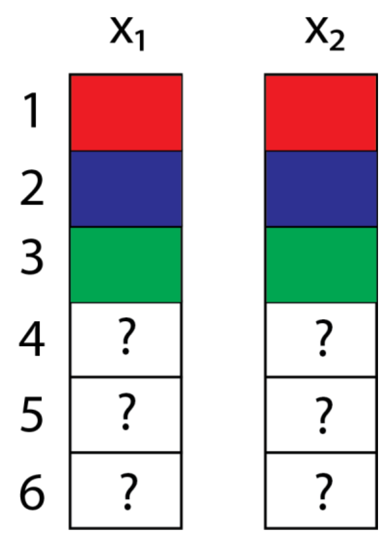
$$\min_{\boldsymbol{\theta}} \quad \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data







$$\mathbb{E}[x_1^T x_2] \approx \frac{1}{6}\sum_{i=1}^{6} x_{1i} x_{2i}$$

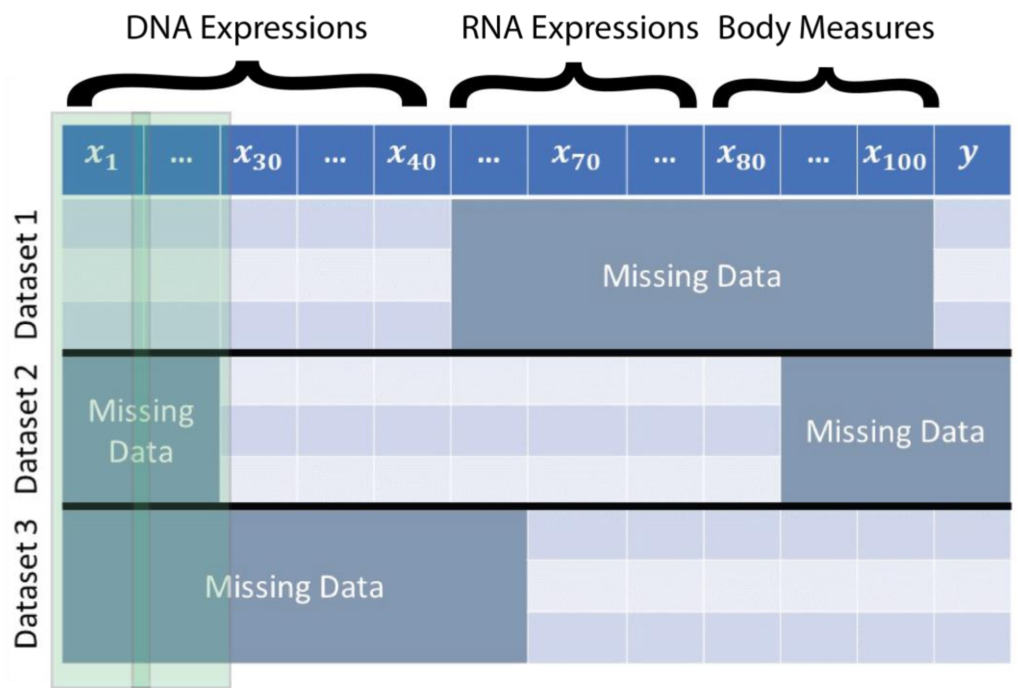
Robust Inference without Imputation

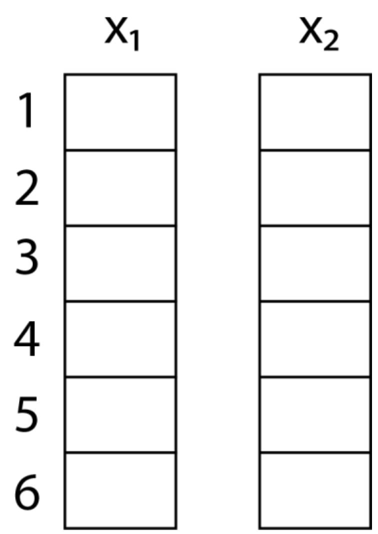
- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

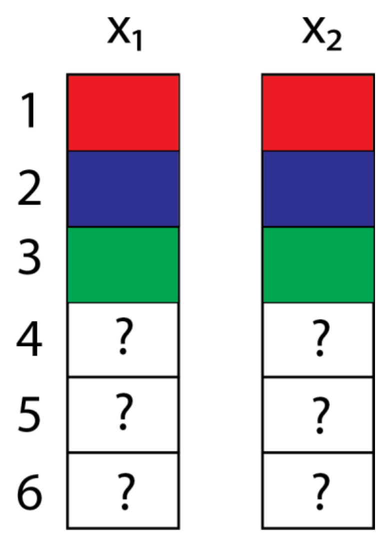
$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^{n}} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data







$$\mathbb{E}[x_1^T x_2] \approx \frac{1}{6}\sum_{i=1}^{6} x_{1i}x_{2i} \qquad \mathbb{E}[x_1^T x_2] \approx \frac{1}{3}(x_{11}x_{21}$$

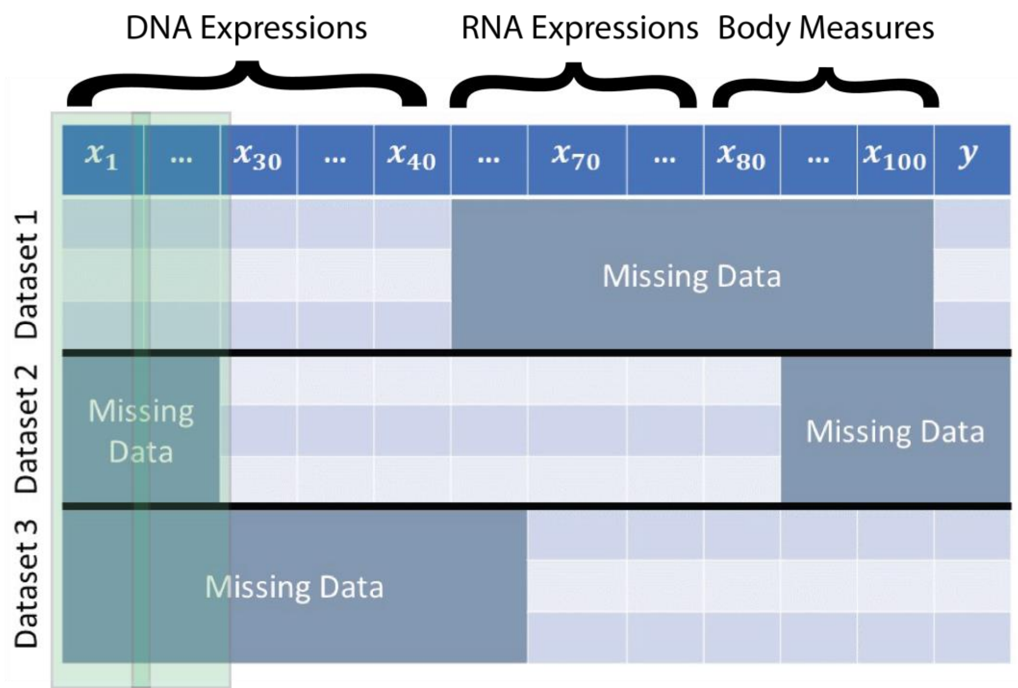
Robust Inference without Imputation

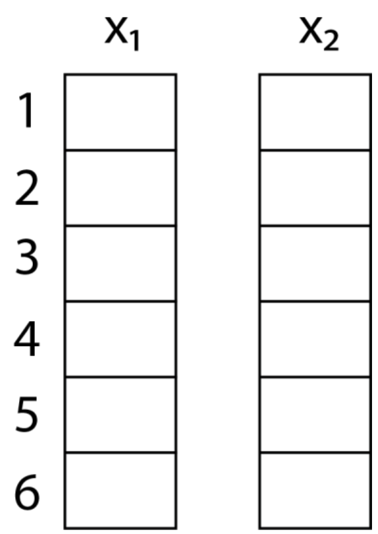
- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

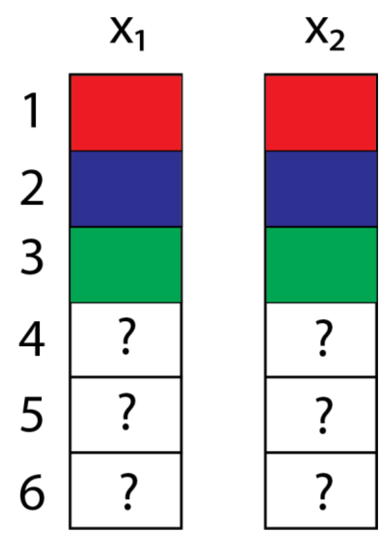
$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data







$$\mathbb{E}[x_1^T x_2] \approx \frac{1}{6} \sum_{i=1}^{6} x_{1i} x_{2i} \qquad \mathbb{E}[x_1^T x_2] \approx \frac{1}{3}(x_{11}x_{21} + x_{12}x_{22})$$

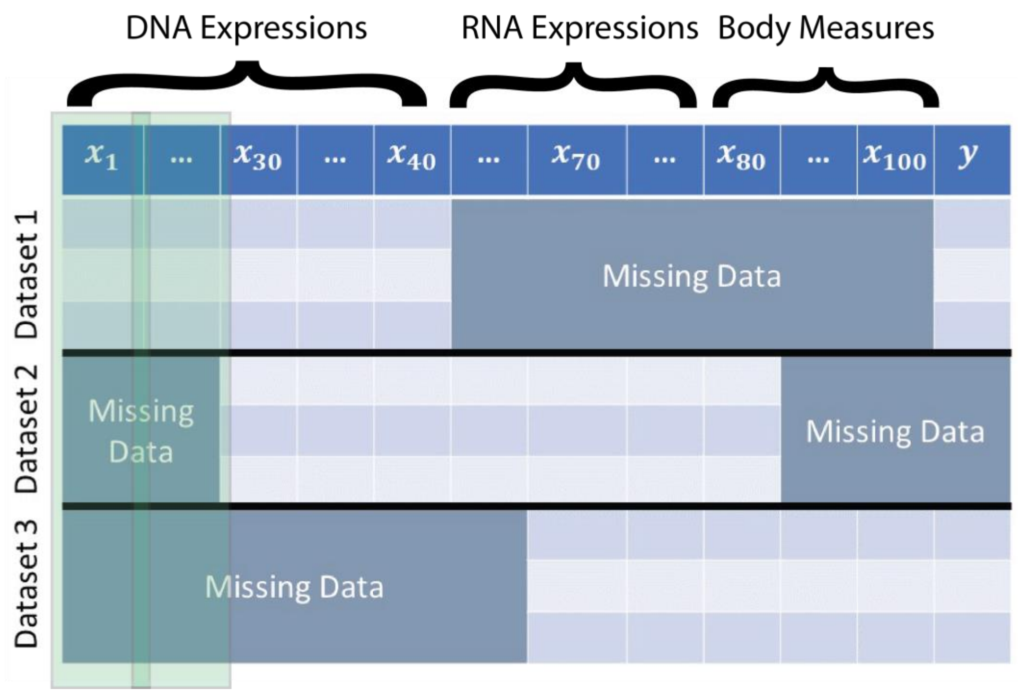
Robust Inference without Imputation

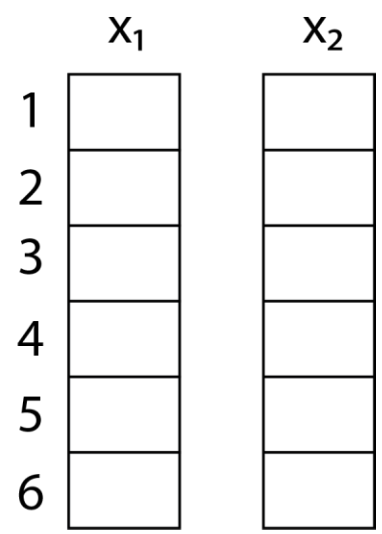
- **Prior Work:** robustness over uncertainty sets around data points [Xu et al., 2009]

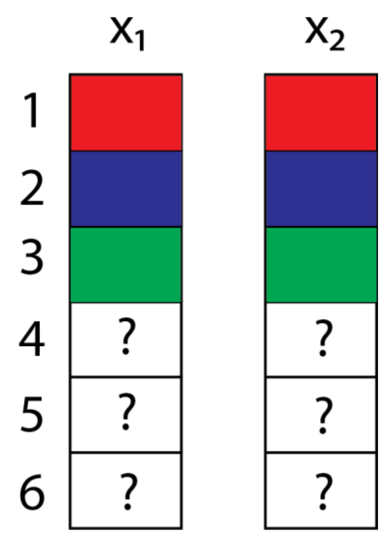
$$\min_{\boldsymbol{\theta}} \max_{\{\boldsymbol{\delta}_i \in \mathcal{N}_i\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i - \boldsymbol{\delta}_i, y_i; \boldsymbol{\theta})$$

  - Too many hyper-parameters (one per data point!)

- Our Idea: Estimating first (mean) and second-order (covariance matrix) moments of the data distribution based on the available data







$$\mathbb{E}[x_1^T x_2] \approx \frac{1}{6}\sum_{i=1}^{6} x_{1i}x_{2i} \qquad \mathbb{E}[x_1^T x_2] \approx \frac{1}{3}(x_{11}x_{21} + x_{12}x_{22} + x_{13}x_{23})$$

RIFLE: Robust Inference without Imputation

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\ \text{s.t.} \quad & \mathbb{E}_P[\mathbf{z}] = \hat{\boldsymbol{\mu}}, \\ & \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] = \hat{C}. \end{aligned}$$

RIFLE: Robust Inference without Imputation

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\ \text{s.t.} \quad & \mathbb{E}_P[\mathbf{z}] = \hat{\boldsymbol{\mu}}, \\ & \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] = \hat{C}. \end{aligned}$$

- Estimations can be inaccurate for low-sample, high-dimensional, and/or datasets with a large proportion of missing values

RIFLE: Robust Inference without Imputation

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\ \text{s.t.} \quad & \mathbb{E}_P[\mathbf{z}] = \hat{\boldsymbol{\mu}}, \\ & \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] = \hat{C}. \end{aligned}$$

- Estimations can be inaccurate for low-sample, high-dimensional, and/or datasets with a large proportion of missing values

- Estimating confidence intervals for first and second order moments using **bootstrap**

RIFLE: Robust Inference without Imputation

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\ \text{s.t.} \quad & \mathbb{E}_P[\mathbf{z}] = \hat{\boldsymbol{\mu}}, \\ & \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] = \hat{C}. \end{aligned}$$

- Estimations can be inaccurate for low-sample, high-dimensional, and/or datasets with a large proportion of missing values

- Estimating confidence intervals for first and second order moments using **bootstrap**

- Solving a distributionally robust optimization over estimated confidence intervals

RIFLE: Robust Inference without Imputation

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\
\text{s.t.} \quad & \mathbb{E}_P[\mathbf{z}] = \hat{\boldsymbol{\mu}}, \\
& \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] = \hat{C}.
\end{aligned}
$$

- Estimations can be inaccurate for low-sample, high-dimensional, and/or datasets with a large proportion of missing values

- Estimating confidence intervals for first and second order moments using **bootstrap**

- Solving a distributionally robust optimization over estimated confidence intervals

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\
\text{s.t.} \quad & \boldsymbol{\mu}_{\min} \leq \mathbb{E}_P[\mathbf{z}] \leq \boldsymbol{\mu}_{\max}, \\
& \mathbf{C}_{\min} \leq \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] \leq \mathbf{C}_{\max}.
\end{aligned}
$$

RIFLE: Robust Inference without Imputation

$$\begin{aligned}\min_{\boldsymbol{\theta}} \quad & \max_{P} \ \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\ & \text{s.t.} \quad \mathbb{E}_P[\mathbf{z}] = \hat{\boldsymbol{\mu}}, \\ & \qquad\quad \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] = \hat{C}.\end{aligned}$$

- Estimations can be inaccurate for low-sample, high-dimensional, and/or datasets with a large proportion of missing values
- Estimating confidence intervals for first and second order moments using **bootstrap**
- Solving a distributionally robust optimization over estimated confidence intervals

$$\begin{aligned}\min_{\boldsymbol{\theta}} \quad & \max_{P} \ \mathbb{E}_P[\ell(\mathbf{z};\boldsymbol{\theta})] \\ & \text{s.t.} \quad \boldsymbol{\mu}_{\min} \leq \mathbb{E}_P[\mathbf{z}] \leq \boldsymbol{\mu}_{\max}, \\ & \qquad\quad \mathbf{C}_{\min} \leq \mathbb{E}_P[\mathbf{z}\mathbf{z}^T] \leq \mathbf{C}_{\max}.\end{aligned}$$

- The proposed min-max problem is intractable in general.

Distributionally Robust Ridge Regression

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[(\boldsymbol{\theta}^T\mathbf{x} - y)^2] + \lambda\|\boldsymbol{\theta}\|_2^2 \\
\text{s.t.} \quad & \boldsymbol{\mu}_{\min} \leq \mathbb{E}_P[(\mathbf{x}, y)] \leq \boldsymbol{\mu}_{\max} \\
& \mathbf{C}_{\min} \leq \mathbb{E}_P[(\mathbf{x}, y)(\mathbf{x}, y)^T] \leq \mathbf{C}_{\max}
\end{aligned}
$$

Distributionally Robust Ridge Regression

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[(\boldsymbol{\theta}^T\mathbf{x} - y)^2] + \lambda\|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\mu}_{\min} \leq \mathbb{E}_P[(\mathbf{x}, y)] \leq \boldsymbol{\mu}_{\max} \\ & \mathbf{C}_{\min} \leq \mathbb{E}_P[(\mathbf{x}, y)(\mathbf{x}, y)^T] \leq \mathbf{C}_{\max} \end{aligned}$$

- Expanding the objective function leads to:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T\mathbf{C}\boldsymbol{\theta} - 2\mathbf{b}^T\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\ & \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\ & \mathbf{C} \succeq 0 \end{aligned}$$

Distributionally Robust Ridge Regression

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{P} \quad & \mathbb{E}_P[(\boldsymbol{\theta}^T\mathbf{x} - y)^2] + \lambda\|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\mu}_{\min} \leq \mathbb{E}_P[(\mathbf{x}, y)] \leq \boldsymbol{\mu}_{\max} \\ & \mathbf{C}_{\min} \leq \mathbb{E}_P[(\mathbf{x}, y)(\mathbf{x}, y)^T] \leq \mathbf{C}_{\max} \end{aligned}$$

- Expanding the objective function leads to:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T\mathbf{C}\boldsymbol{\theta} - 2\mathbf{b}^T\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\ & \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\ & \mathbf{C} \succeq 0 \end{aligned}$$

- How can we solve this problem efficiently?

First Idea for Solving the Robust Ridge Regression

$$
\min_{\boldsymbol{\theta}} \quad \overbrace{\begin{aligned} \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\ & \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\ & \mathbf{C} \succeq 0 \end{aligned}}^{g(\boldsymbol{\theta})}
$$

First Idea for Solving the Robust Ridge Regression

$$
\min_{\boldsymbol{\theta}} \; \overbrace{\begin{aligned} \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \le \mathbf{C} \le \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\ & \hat{\mathbf{b}} - \boldsymbol{\delta} \le \mathbf{b} \le \hat{\mathbf{b}} + \boldsymbol{\delta}, \\ & \mathbf{C} \succeq 0 \end{aligned}}^{g(\boldsymbol{\theta})}
$$

- Using Danskin's theorem, applying gradient descent to $g(\boldsymbol{\theta})$

First Idea for Solving the Robust Ridge Regression

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad & \overbrace{\max_{\mathbf{C},\mathbf{b}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}^{g(\boldsymbol{\theta})} \\
& \quad \text{s.t.} \quad \hat{\mathbf{C}} - \boldsymbol{\Delta} \le \mathbf{C} \le \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \qquad\qquad \hat{\mathbf{b}} - \boldsymbol{\delta} \le \mathbf{b} \le \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \qquad\qquad \mathbf{C} \succeq 0
\end{aligned}
$$

- Using Danskin's theorem, applying gradient descent to $g(\boldsymbol{\theta})$

- No closed-form for $g(\boldsymbol{\theta})$

First Idea for Solving the Robust Ridge Regression

$$
\min_{\boldsymbol{\theta}} \; \underbrace{\max_{\mathbf{C},\mathbf{b}} \; \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}_{g(\boldsymbol{\theta})}
$$

$$
\begin{aligned}
\text{s.t.} \quad & \boxed{\hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta},} \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \boxed{\mathbf{C} \succeq 0}
\end{aligned}
$$

- Using Danskin's theorem, applying gradient descent to $g(\boldsymbol{\theta})$
- No closed-form for $g(\boldsymbol{\theta})$

First Idea for Solving the Robust Ridge Regression

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad & \overbrace{\max_{\mathbf{C},\mathbf{b}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}^{g(\boldsymbol{\theta})} \\
& \text{s.t.} \quad \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \qquad \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \qquad \mathbf{C} \succeq 0
\end{aligned}
$$

- Using Danskin's theorem, applying gradient descent to $g(\boldsymbol{\theta})$
- No closed-form for $g(\boldsymbol{\theta})$
- **Observation**: the problem is convex-concave with convex constraints

Finding an Efficient Practical Solver

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

Finding an Efficient Practical Solver

$$\min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad \boldsymbol{\theta}^T\mathbf{C}\boldsymbol{\theta} - 2\mathbf{b}^T\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_2^2$$

$$\text{s.t.} \quad \hat{\mathbf{C}} - \boldsymbol{\Delta} \le \mathbf{C} \le \hat{\mathbf{C}} + \boldsymbol{\Delta},$$

$$\hat{\mathbf{b}} - \boldsymbol{\delta} \le \mathbf{b} \le \hat{\mathbf{b}} + \boldsymbol{\delta},$$

$$\mathbf{C} \succeq 0$$

Finding an Efficient Practical Solver

$$\begin{aligned} \max_{\mathbf{C},\mathbf{b}} \quad \min_{\boldsymbol{\theta}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\ & \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\ & \mathbf{C} \succeq 0 \end{aligned}$$

Finding an Efficient Practical Solver

$$
\begin{aligned}
\max_{\mathbf{C},\mathbf{b}} \quad & \overbrace{\min_{\boldsymbol{\theta}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}^{g(\mathbf{C},\mathbf{b})} \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

Finding an Efficient Practical Solver

$$
\begin{aligned}
\max_{\mathbf{C},\mathbf{b}} \quad & \overbrace{\min_{\boldsymbol{\theta}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}^{g(\mathbf{C},\mathbf{b})} \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

- The minimization problem has a closed-form solution

Finding an Efficient Practical Solver

$$
\begin{aligned}
\max_{\mathbf{C},\mathbf{b}} \quad & \overbrace{\min_{\boldsymbol{\theta}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}^{g(\mathbf{C},\mathbf{b})} \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \le \mathbf{C} \le \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \le \mathbf{b} \le \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

- The minimization problem has a closed-form solution
- Applying projected gradient ascent on $g(\mathbf{C},\mathbf{b})$ leads to:

Finding an Efficient Practical Solver

$$\max_{\mathbf{C},\mathbf{b}} \quad \overbrace{\min_{\boldsymbol{\theta}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}^{g(\mathbf{C},\mathbf{b})}$$

$$\text{s.t.} \quad \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta},$$

$$\hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta},$$

$$\mathbf{C} \succeq 0$$

- The minimization problem has a closed-form solution
- Applying projected gradient ascent on $g(\mathbf{C}, \mathbf{b})$ leads to:

Algorithm Projected Gradient Ascent on Robust Ridge Regression

1: **for** $i = 1, \ldots, T$ **do**
2: Update $\mathbf{C} = [\Pi_{\boldsymbol{\Delta}}(\mathbf{C} + \alpha \boldsymbol{\theta}\boldsymbol{\theta}^T)]_+$
3: Update $\mathbf{b} = \Pi_{\delta}(\mathbf{b} - 2\alpha\boldsymbol{\theta})$
4: Set $\boldsymbol{\theta} = (\mathbf{C} + \lambda \mathbf{I})^{-1}\mathbf{b}$
5: **end for**

Finding an Efficient Practical Solver

$$\max_{\mathbf{C},\mathbf{b}} \ \underbrace{\min_{\boldsymbol{\theta}} \quad \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2}_{g(\mathbf{C},\mathbf{b})}$$

$$\begin{aligned} \text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\ & \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\ & \mathbf{C} \succeq 0 \end{aligned}$$

- The minimization problem has a closed-form solution
- Applying projected gradient ascent on $g(\mathbf{C}, \mathbf{b})$ leads to:

Algorithm Projected Gradient Ascent on Robust Ridge Regression

1: **for** $i = 1, \ldots, T$ **do**
2: Update $\boxed{\mathbf{C} = [\Pi_{\boldsymbol{\Delta}}(\mathbf{C} + \alpha \boldsymbol{\theta}\boldsymbol{\theta}^T)]_+}$
3: Update $\mathbf{b} = \Pi_{\delta}(\mathbf{b} - 2\alpha\boldsymbol{\theta})$
4: Set $\boldsymbol{\theta} = (\mathbf{C} + \lambda \mathbf{I})^{-1}\mathbf{b}$
5: **end for**

How to perform both projections?

An Alternative Approach for Projected Gradient Ascent

- How to handle the joint projection to the set of **box constraints** and the **PSD cone**?

An Alternative Approach for Projected Gradient Ascent

- How to handle the joint projection to the set of **box constraints** and the **PSD cone**?
    - Removing the PSD constraint in the implementation (relaxation)

An Alternative Approach for Projected Gradient Ascent

- How to handle the joint projection to the set of **box constraints** and the **PSD cone**?
  - Removing the PSD constraint in the implementation (relaxation)
  - **Idea**: Using the **dual formulation** on the inner maximization problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

An Alternative Approach for Projected Gradient Ascent

- How to handle the joint projection to the set of **box constraints** and the **PSD cone**?

  - Removing the PSD constraint in the implementation (relaxation)

  - **Idea**: Using the **dual formulation** on the inner maximization problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

Writing the dual of inner maximization problem

An Alternative Approach for Projected Gradient Ascent

- How to handle the joint projection to the set of **box constraints** and the **PSD cone**?
  - Removing the PSD constraint in the implementation (relaxation)
  - **Idea**: Using the **dual formulation** on the inner maximization problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad \max_{\mathbf{C},\mathbf{b}} \quad & \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta} - 2\mathbf{b}^T \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_2^2 \\
\text{s.t.} \quad & \hat{\mathbf{C}} - \boldsymbol{\Delta} \leq \mathbf{C} \leq \hat{\mathbf{C}} + \boldsymbol{\Delta}, \\
& \hat{\mathbf{b}} - \boldsymbol{\delta} \leq \mathbf{b} \leq \hat{\mathbf{b}} + \boldsymbol{\delta}, \\
& \mathbf{C} \succeq 0
\end{aligned}
$$

Writing the dual of inner maximization problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d} \rangle + \langle \mathbf{b}_{\max}, \mathbf{e} \rangle - \langle \mathbf{C}_{\min}, \mathbf{A} \rangle + \langle \mathbf{C}_{\max}, \mathbf{B} \rangle + \lambda \|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & -\boldsymbol{\theta}\boldsymbol{\theta}^T - \mathbf{A} + \mathbf{B} - \mathbf{H} = 0, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A}, \mathbf{B}, \mathbf{d}, \mathbf{e} \geq 0, \\
& \mathbf{H} \succeq 0
\end{aligned}
$$

Change of Variables

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d} \rangle + \langle \mathbf{b}_{\max}, \mathbf{e} \rangle - \langle \mathbf{C}_{\min}, \mathbf{A} \rangle + \langle \mathbf{C}_{\max}, \mathbf{B} \rangle + \lambda \|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \boxed{-\boldsymbol{\theta}\boldsymbol{\theta}^T - \mathbf{A} + \mathbf{B} - \mathbf{H} = 0,} \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A}, \mathbf{B}, \mathbf{d}, \mathbf{e} \geq 0, \\
& \mathbf{H} \succeq 0
\end{aligned}
$$

Change of Variables

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \boxed{-\boldsymbol{\theta}\boldsymbol{\theta}^T - \mathbf{A} + \mathbf{B} - \mathbf{H} = 0,} \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A}, \mathbf{B}, \mathbf{d}, \mathbf{e} \geq 0, \\
& \mathbf{H} \succeq 0
\end{aligned}
$$

$$\Big\downarrow \; \boldsymbol{G} = \boldsymbol{H} + \boldsymbol{\theta}\boldsymbol{\theta}^T$$

Change of Variables

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \boxed{-\boldsymbol{\theta}\boldsymbol{\theta}^T - \mathbf{A} + \mathbf{B} - \mathbf{H} = 0,} \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A}, \mathbf{B}, \mathbf{d}, \mathbf{e} \geq 0, \\
& \mathbf{H} \succeq 0
\end{aligned}
$$

$$\Big\downarrow \; \boldsymbol{G} = \boldsymbol{H} + \boldsymbol{\theta}\boldsymbol{\theta}^T$$

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \mathbf{B} - \mathbf{A} = \mathbf{G}, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A}, \mathbf{B}, \mathbf{d}, \mathbf{e} \geq 0, \\
& \mathbf{G} \succeq \boldsymbol{\theta}\boldsymbol{\theta}^T
\end{aligned}
$$

Applying ADMM to the Dual Problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \mathbf{B} - \mathbf{A} = \mathbf{G}, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A} = \mathbf{A}', \mathbf{B} = \mathbf{B}', \\
& \mathbf{d} = \mathbf{d}', \mathbf{e} = \mathbf{e}', \boldsymbol{\theta} = \boldsymbol{\theta}', \\
& \mathbf{A}', \mathbf{B}', \mathbf{d}', \mathbf{e}' \geq 0, \\
& \mathbf{G} \succeq \boldsymbol{\theta}'\boldsymbol{\theta}'^{T}
\end{aligned}
$$

Applying ADMM to the Dual Problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \mathbf{B} - \mathbf{A} = \mathbf{G}, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A} = \mathbf{A}^{'}, \mathbf{B} = \mathbf{B}^{'}, \\
& \mathbf{d} = \mathbf{d}^{'}, \mathbf{e} = \mathbf{e}^{'}, \boldsymbol{\theta} = \boldsymbol{\theta}^{'}, \\
& \mathbf{A}^{'}, \mathbf{B}^{'}, \mathbf{d}^{'}, \mathbf{e}^{'} \geq 0, \\
& \mathbf{G} \succeq \boldsymbol{\theta}^{'}\boldsymbol{\theta}^{'T}
\end{aligned}
$$

- Defining two blocks of variables

Applying ADMM to the Dual Problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \mathbf{B} - \mathbf{A} = \mathbf{G}, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A} = \mathbf{A}', \mathbf{B} = \mathbf{B}', \\
& \mathbf{d} = \mathbf{d}', \mathbf{e} = \mathbf{e}', \boldsymbol{\theta} = \boldsymbol{\theta}', \\
& \mathbf{A}', \mathbf{B}', \mathbf{d}', \mathbf{e}' \geq 0, \\
& \mathbf{G} \succeq \boldsymbol{\theta}'\boldsymbol{\theta}'^T
\end{aligned}
$$

- Defining two blocks of variables

$$\mathbf{w} = (\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{e}, \boldsymbol{G}, \boldsymbol{B}', \boldsymbol{A}')$$

$$\mathbf{z} = (\boldsymbol{\theta}', \boldsymbol{d}', \boldsymbol{e}', \boldsymbol{B}, \boldsymbol{A})$$

Applying ADMM to the Dual Problem

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \mathbf{B} - \mathbf{A} = \mathbf{G}, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A} = \mathbf{A}', \mathbf{B} = \mathbf{B}', \\
& \mathbf{d} = \mathbf{d}', \mathbf{e} = \mathbf{e}', \boldsymbol{\theta} = \boldsymbol{\theta}', \\
& \mathbf{A}', \mathbf{B}', \mathbf{d}', \mathbf{e}' \geq 0, \\
& \mathbf{G} \succeq \boldsymbol{\theta}'\boldsymbol{\theta}'^T
\end{aligned}
$$

- Defining two blocks of variables

$$\mathbf{w} = (\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{e}, \boldsymbol{G}, \boldsymbol{B}', \boldsymbol{A}')$$

$$\mathbf{z} = (\boldsymbol{\theta}', \boldsymbol{d}', \boldsymbol{e}', \boldsymbol{B}, \boldsymbol{A})$$

Algorithm ADMM for Two Blocks

1: **for** $t = 1, \ldots, T$ **do**
2: $\quad \mathbf{w}^{t+1} = \arg\min_{\mathbf{w}} f(\mathbf{w}) + \langle \mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{z}^t - \mathbf{c}, \boldsymbol{\lambda}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{z}^t - \mathbf{c}\|^2$
3: $\quad \mathbf{z}^{t+1} = \arg\min_{\mathbf{z}} f(\mathbf{w}^{t+1}) + \langle \mathbf{A}\mathbf{w}^{t+1} + \mathbf{B}\mathbf{z} - \mathbf{c}, \boldsymbol{\lambda}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{w}^{t+1} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2$
4: $\quad \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \rho(\mathbf{A}\mathbf{w}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{c})$
5: **end for**

Applying ADMM to the Dual Problem

$$\begin{aligned}
\min_{\boldsymbol{\theta},\mathbf{A},\mathbf{B},\mathbf{d},\mathbf{e},\mathbf{H}} \quad & -\langle \mathbf{b}_{\min}, \mathbf{d}\rangle + \langle \mathbf{b}_{\max}, \mathbf{e}\rangle - \langle \mathbf{C}_{\min}, \mathbf{A}\rangle + \langle \mathbf{C}_{\max}, \mathbf{B}\rangle + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \mathbf{B} - \mathbf{A} = \mathbf{G}, \\
& 2\boldsymbol{\theta} - \mathbf{d} + \mathbf{e} = 0, \\
& \mathbf{A} = \mathbf{A}', \mathbf{B} = \mathbf{B}', \\
& \mathbf{d} = \mathbf{d}', \mathbf{e} = \mathbf{e}', \boldsymbol{\theta} = \boldsymbol{\theta}', \\
& \mathbf{A}', \mathbf{B}', \mathbf{d}', \mathbf{e}' \geq 0, \\
& \mathbf{G} \succeq \boldsymbol{\theta}'\boldsymbol{\theta}'^T
\end{aligned}$$

- Defining two blocks of variables

$$\mathbf{w} = (\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{e}, \boldsymbol{G}, \boldsymbol{B}', \boldsymbol{A}')$$

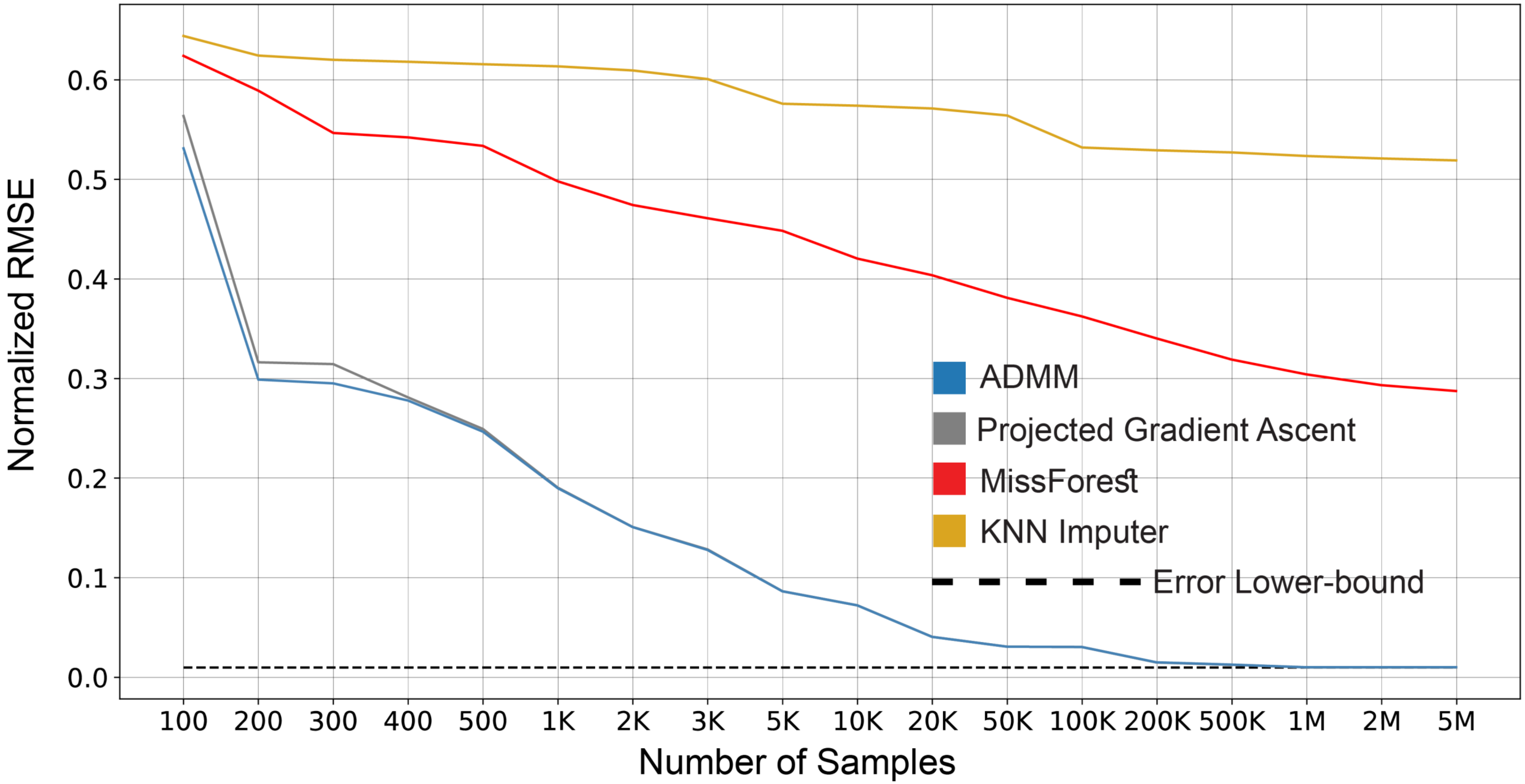
$$\mathbf{z} = (\boldsymbol{\theta}', \boldsymbol{d}', \boldsymbol{e}', \boldsymbol{B}, \boldsymbol{A})$$

Algorithm ADMM for Two Blocks

1: **for** $t = 1, \ldots, T$ **do**
2: $\quad \mathbf{w}^{t+1} = \arg\min_{\mathbf{w}} f(\mathbf{w}) + \langle \mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{z}^t - \mathbf{c}, \boldsymbol{\lambda}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{z}^t - \mathbf{c}\|^2$
3: $\quad \mathbf{z}^{t+1} = \arg\min_{\mathbf{z}} f(\mathbf{w}^{t+1}) + \langle \mathbf{A}\mathbf{w}^{t+1} + \mathbf{B}\mathbf{z} - \mathbf{c}, \boldsymbol{\lambda}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{w}^{t+1} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2$
4: $\quad \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \rho(\mathbf{A}\mathbf{w}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{c})$
5: **end for**

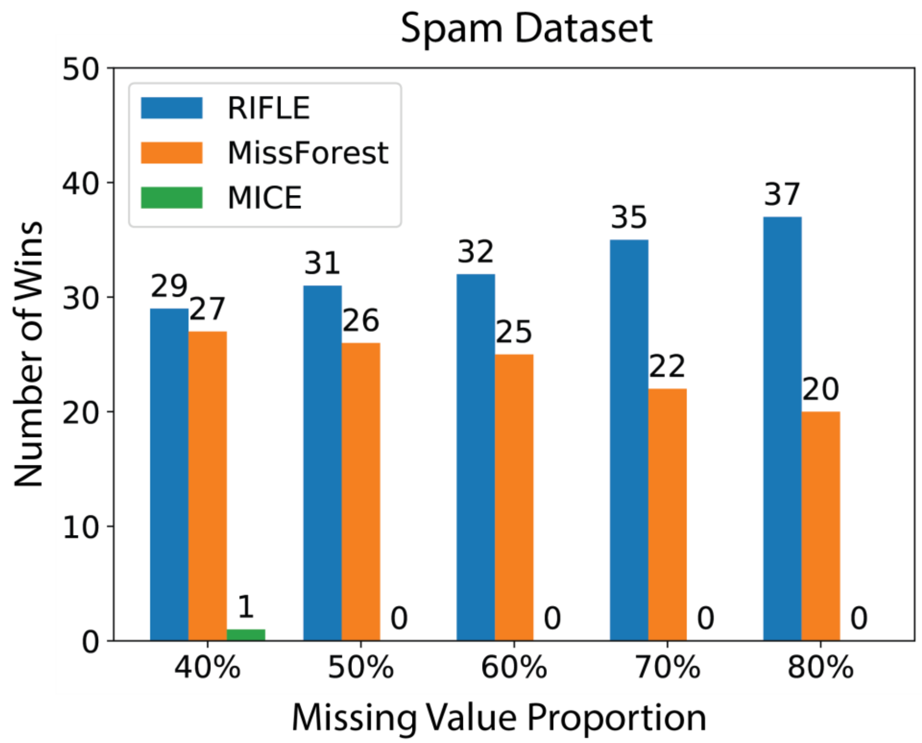
Proposition. *If the feasible set has non-empty interior, then RIFLE converges to an ϵ-optimal solution of the problem in $\mathcal{O}(\frac{1}{\epsilon})$ iterations.*

RIFLE Consistency

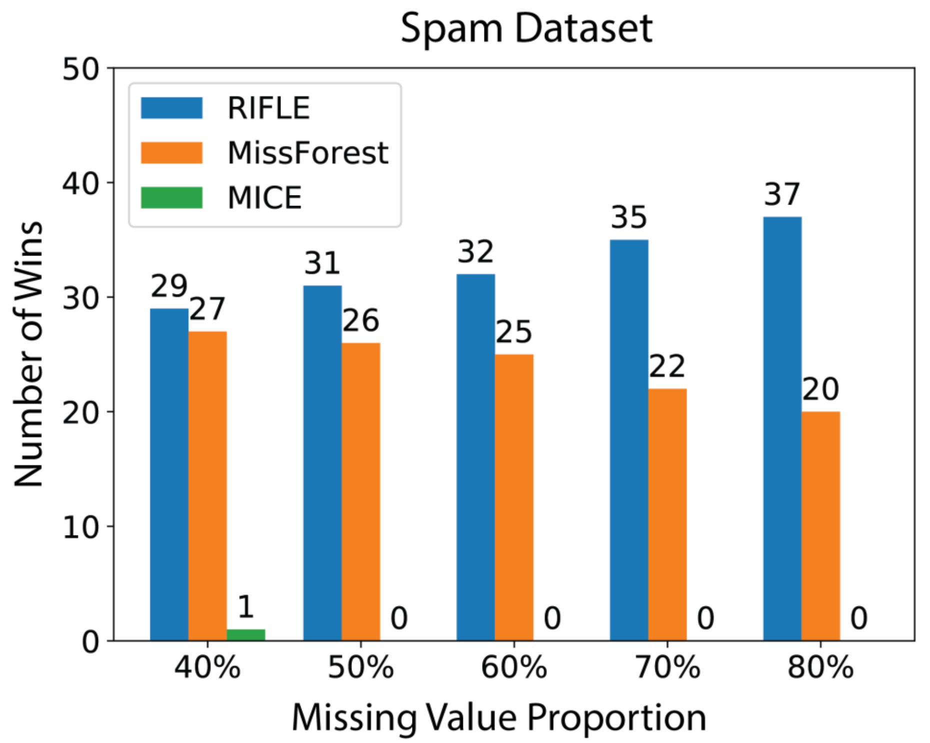


- **Jointly normal** dataset with **linear** relation between the predictors and the target
- **40%** missing values and **100** features
- Changing the number of samples from **100 to 5 Million**

How Many Times RIFLE Outperforms All Existing Packages?

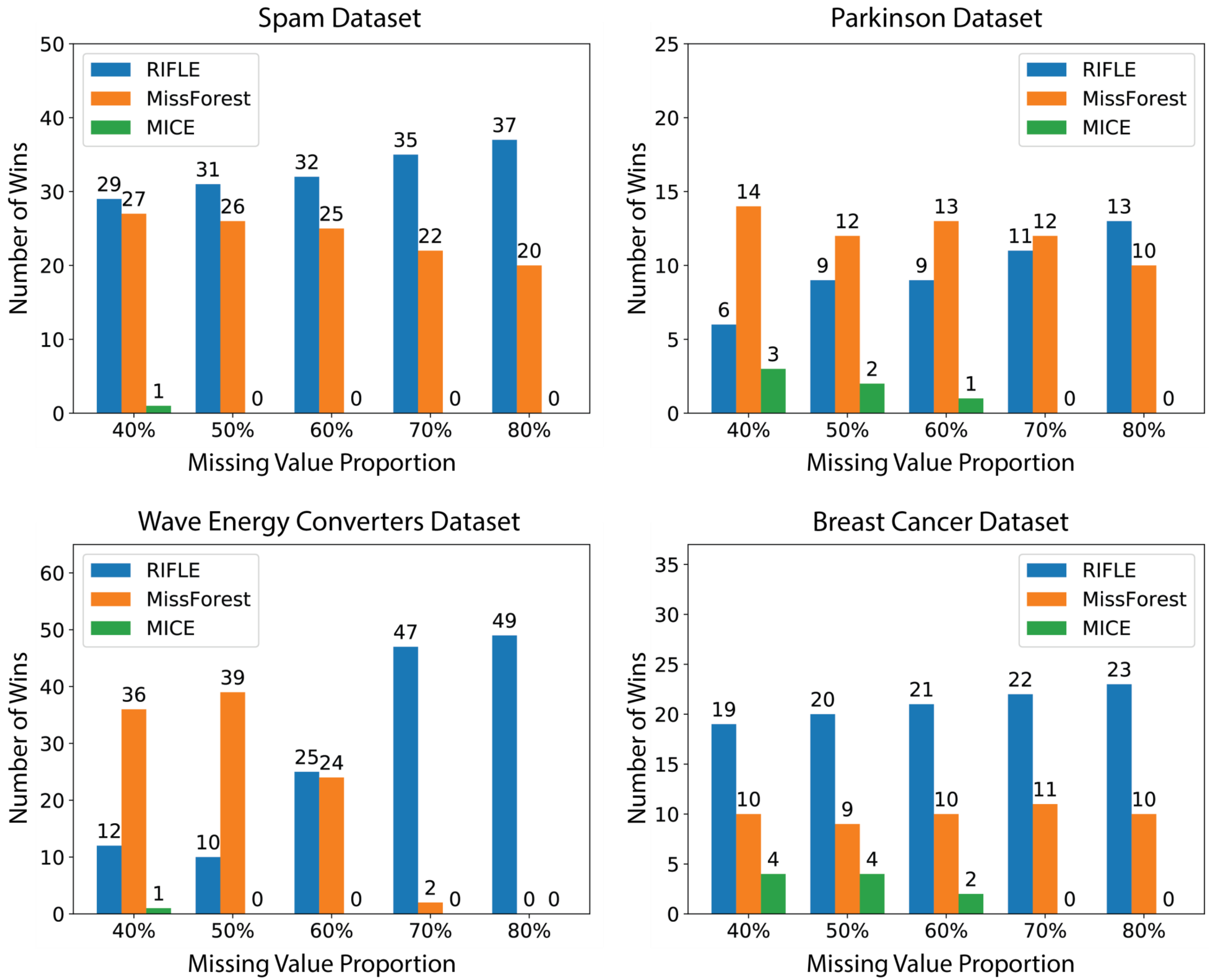


How Many Times RIFLE Outperforms All Existing Packages?



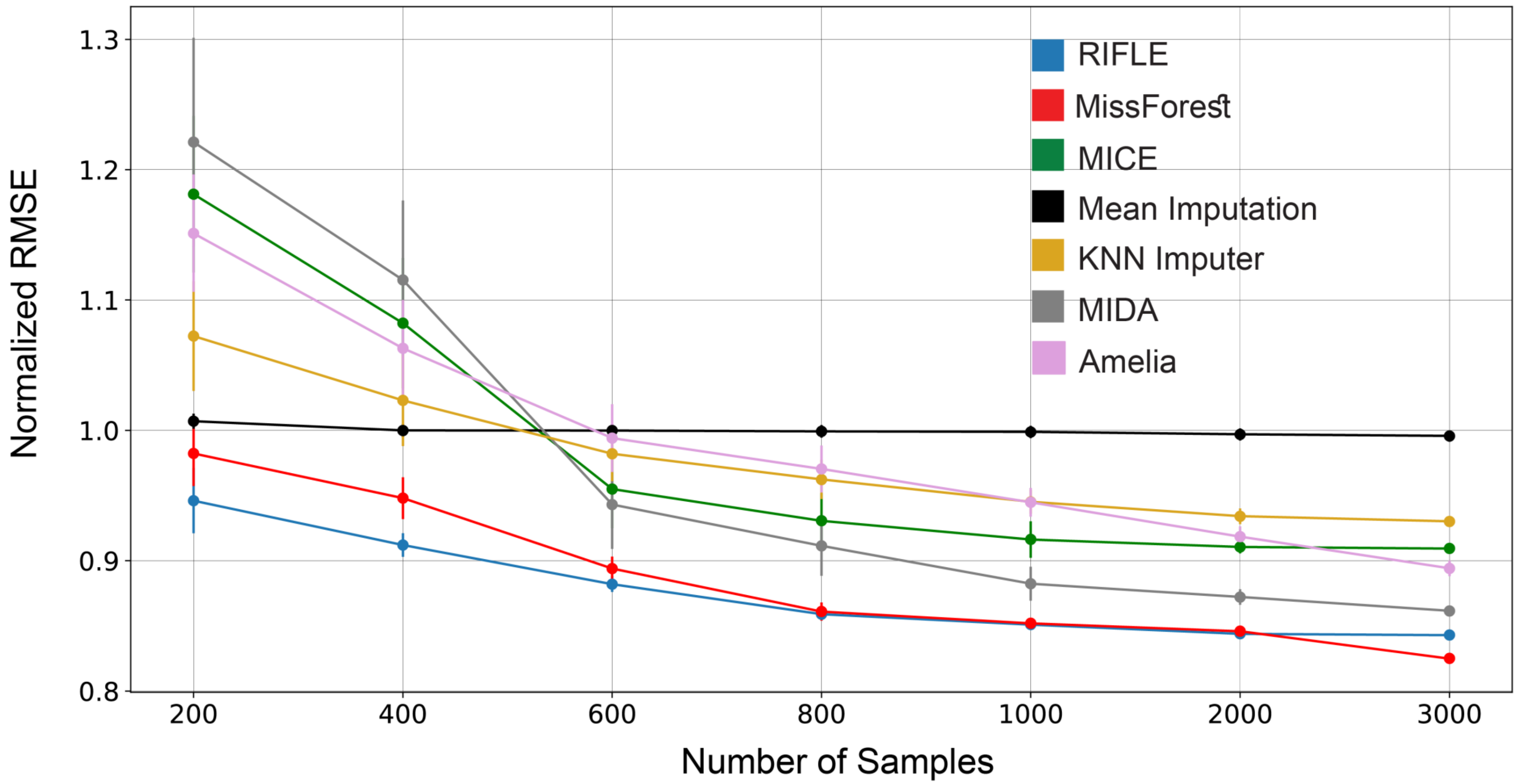
- RIFLE wins more than the best imputer packages

How Many Times RIFLE Outperforms All Existing Packages?



➢ RIFLE wins more than the best imputer packages

RIFLE Outperforms Other Algorithms for Lower Samples



- Evaluation on Drive dataset (40% missing values completely at random)

MissForest: Stekhoven, Daniel J., and Peter Bühlmann. "MissForest: non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28, (2012).

MICE: Royston, Patrick, and Ian R. White. "Multiple imputation by chained equations (MICE): implementation in Stata." *Journal of statistical software* 45 (2011).

Mean Imputer: Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, (2019).

KNN Imputer: Troyanskaya, Olga et al., "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17 (2001).

MIDA: Gondara, Lovedeep, and Ke Wang. "Mida: Multiple imputation using denoising autoencoders." In *PKDD (*2018).

Amelia: Honaker, James, Gary King, and Matthew Blackwell. "Amelia II: A program for missing data." *Journal of statistical software* 45 (2011).

Reference

- **Sina Baharlouei**, Kelechi Ogudu, Peng Dai, Sze-chuan Suen and Meisam Razaviyayn. “RIFLE: Imputation and Robust Inference from Low Order Marginals" In ICML Workshop on Duality Principles for Modern Machine Learning, 2023.
  - RIFLE Package: https://github.com/optimization-for-data-driven-science/RIFLE.